Реализация высокоскоростной сети для суперкомпьютерных систем: проблемы, результаты, развитие

В. А. Михеев, А. С. Симонов, А. И. Слуцкин, Д. В. Макагон, Д. А. Поляков, Е. Л. Сыромятников, И. А. Жабин, А. Н. Щербак, Е. Р. Куштанов, А. Е. Леонова

4 ноября 2013 г.

Аннотация

Реализация высокоскоростной коммуникационной сети на многих задачах в значительной степени определяет реальную производительность суперкомпьютерной системы. Разработка эффективной высокоскоростной коммуникационной сети является сложной научно-технической задачей, требующей совместной работы высококвалифицированных специалистов из многих областей. В докладе рассказывается об основных вопросах, требующих решения при разработке архитектуры сети и программного обеспечения. Акцент сделан на необходимости принятия решений по многим ключевым параметрам, включая топологию сети, алгоритмы маршрутизации, протоколы передачи данных, поддерживаемые модели параллельного программирования, форм-фактор сетевых адаптеров и кабелей и т. д. Также в докладе кратко описано современное состояние отрасли разработки высокоскоростных сетей для суперкомпьютеров и сформулированы некоторые мировые тенденции в данной области.

Введение

Наблюдаемый последние десятилетия неуклонный рост мощности суперкомпьютерных систем во многом обусловлен постоянным увеличением числа узлов, процессоров, ядер и внедрением различных ускорителей, при этом эффективность использования суперкомпьютерных ресурсов (памяти, процессорного времени, ускорителей) и достигаемая производительность на многих задачах в значительной степени определяются реализацией высокоскоростной коммуникационной сети, которая обеспечивает обмен данными и синхронизацию вычислительных узлов.

Наиболее мощными суперкомпьютерами на текущий момент (согласно списку Top500, июнь 2013) являются: китайские системы Tianhe-2 и Tianhe-1A, японский K Computer, американские Cray Titan, IBM Blue Gene/Q. Все эти суперкомпьютеры используют собственные уникальные («заказные») коммуникационные сети, разрабатываемые в составе этих вычислительных систем, и доступные только совместно с ними. Приобретение подобных машин в России в ряде случаев затруднено, а зачастую является фактически невозможным. В то же время коммерчески доступные сети InfiniBand и Ethernet далеко не всегда подходят для эффективной реализации систем со столь высокими требованиями по масштабируемости, надёжности и производительности. В связи с этим крайне актуальным является вопрос разработки отечественной высокоскоростной сети, сравнимой с западными «заказными» аналогами.

В ОАО «НИЦЭВТ» с 2006-го года ведётся разработка коммуникационной сети «Ангара» — отечественной высокоскоростной коммуникационной сети с топологией 4D-тор, которая сможет стать основой для создания отечественных суперкомпьютеров [1, 2]. В 2013 году, став результатом длительной подготовительной работы, появилось на свет первое поколение маршрутизаторов сети «Ангара» на базе СБИС ЕС8430. В процессе разработки коммуникационной сети перед инженерами встал целый ряд вопросов, требующих решений, и предполагающих расстановку приоритетов между ценой, производительностью, эффективностью энергопотребления и другими требованиями, во многом конфликтующими между собой, поскольку часто попытки улучшения одной характеристики могут приводить к ухудшению другой.

Далее рассматриваются основные вопросы, которые решались при разработке архитектуры и программного обеспечения для сети «Ангара», даётся обоснование решений, принятых по многим ключевым вопросам, включая топологию сети, алгоритмы маршрутизации, протоколы передачи

данных, поддерживаемые модели параллельного программирования, форм-фактор сетевых адаптеров и кабелей и т. д.

1. Высокоскоростные сети

Коммуникационная сеть состоит из узлов, в каждом из которых есть сетевой адаптер, соединенный с одним или несколькими маршрутизаторами, которые в свою очередь соединяются между собой высокоскоростными каналами связи (линками) [1]. Структура сети, определяющая, как именно связаны между собой узлы системы, задается топологией сети. В настоящее время распространены топологии многомерный тор, fat tree, dragonfly.

Архитектура маршрутизатора определяет структуру и функциональность блоков, отвечающих за передачу данных между узлами сети, а также необходимые свойства протоколов канального, сетевого и транспортного уровней, включая алгоритмы маршрутизации, арбитража и управления потоком данных. Архитектура сетевого адаптера определяет структуру и функциональность блоков, отвечающих за взаимодействие между процессором, памятью и сетью; в частности, на этом уровне осуществляется поддержка MPI-операций, RDMA (Remote Direct Memory Access — прямой доступ к памяти другого узла без участия его процессора), подтверждений получения другим узлом пакета, обработки исключительных ситуаций, агрегации пакетов.

Для оценки производительности коммуникационной сети чаще всего используются три характеристики: пропускная способность, коммуникационная задержка, темп выдачи сообщений. Для полноты картины данные характеристики измеряются на разных видах трафика, например, когда один узел рассылает данные всем остальным, либо, наоборот, все узлы шлют данные одному, либо когда все узлы посылают данные случайным адресатам.

Если посмотреть на статистику списка Top500, то можно выяснить, что большинство представленных в нём систем используют коммерчески доступные сети InfiniBand и Ethernet. Сеть InfiniBand широко используется для построения кластерных систем и суперкомпьютеров. Последнее на данный момент поколение сети InfiniBand — InfiniBand FDR — было представлено в июне 2011 года. Основным количественным улучшением нового поколения является увеличенная пропускная способность линков — до 14 Гбит/с. Существующие реализации архитектуры сети Inifini-Band оптимизированы под топологию fat tree, однако последние поколения коммутаторов и маршрутизаторов поддерживают топологию многомерный тор, а также гибридную топологию из fat tree и трёхмерного тора. Сеть Ethernet традиционно занимает нишу, где обмен данными между узлами некритичен.

В отличие от коммерчески доступных сетей, «заказные» сети занимают гораздо меньшую долю рынка, однако именно они используются в наиболее мощных суперкомпьютерах.

Китайский суперкомпьютер Tianhe-1A состоит из 7168 вычислительных узлов, объединенных сетью Arch с топологией fat tree. Сеть строится из 16-портовых маршрутизаторов, односторонняя пропускная способность линка — 8 ГБ/с, задержка — 1,57 мкс. В суперкомпьютере Tianhe-2 используется сеть TH Express-2 с топологией fat tree. На верхнем уровне сети используются 13 576-портовых коммутаторов на базе специально разработанного чипа NRC, агрегатная пропускная способность которого составляет 2,56 Тбит/с. Коммуникационная задержка для этой сети, измеренная на сообщениях размером 1 КБ на 12000 узлах, равна 9 мкс [3].

Системы серии IBM Blue Gene являются классическими представителями суперкомпьютеров, использующих топологию многомерный тор для объединения вычислительных узлов. В первых двух поколениях этих систем — Blue Gene/L (2004) и Blue Gene/P (2007) — использовалась топология 3D-тор, дополненная рядом специализированных сетей для синхронизации и коллективных операций; в Blue Gene/Q (2012) реализована топология 5D-тор без дополнительных сетей. Пропускная способность линка в Blue Gene/Q составляет 2 ГБ/с, что, с одной стороны, существенно больше 0,425 ГБ/с, предоставляемых в предыдущем поколении, но с другой — на порядок меньше пропускной способности, предоставляемой, например, в сетях InfiniBand или Cray Gemini.

Сеть Tofu (от Torus Fusion), которая используется в японском суперкомпьютере K Computer, имеет топологию многомерный тор. Узел сети Tofu имеет 10 линков с пропускной способностью в 40 Гбит/с каждый.

Ряд отечественных организаций также ведёт разработку коммуникационных сетей для использования в суперкомпьютерах, в том числе РФЯЦ ВНИИЭФ, Институт программных систем РАН и РСК «СКИФ», ИПМ РАН и НИИ «Квант» (сеть «МВС-Экспресс»).

2. Разработка высокоскоростной сети «Ангара»: ключевые вопросы

Высокоскоростная коммуникационная сеть «Ангара», разрабатываемая в ОАО «НИЦЭВТ», имеет топологию 4D-тор. Основной целью при разработке сети является создание отечественной «заказной» сети, сравнимой с мировыми аналогами, которая может использоваться в суперкомпьютерах вплоть до транспетафлопсного уровня производительности. СБИС ЕС8430 (рис. 1) является основой первого поколения маршрутизаторов сети «Ангара».



Рисунок 1. СБИС «Ангара» (ЕС8430)

Началом разработки сети «Ангара» стала проведённая в 2006 году совокупность работ по имитационному моделированию различных вариантов сети и изучению основных решений по топологии, архитектуре маршрутизатора, алгоритмам маршрутизации и арбитражу. Изначально помимо тороидальной топологии рассматривались сети Кэли и fat tree. Четырехмерный тор был выбран в силу более простой маршрутизации, хорошей масштабируемости, высокой связности по сравнению с торами меньшей размерности. Моделирование сети позволило изучить влияние различных параметров архитектуры сети на основные характеристики производительности, понять некоторые закономерности для трафика задач с интенсивным нерегулярным доступом к памяти. В результате были подобраны различные количественные характеристики будущего маршрутизатора, такие как оптимальные размеры буферов и число виртуальных каналов; были проанализированы потенциальные узкие места. При разработке принципов работы сети в качестве руководства использовались [4] и [5], некоторые идеи были также в том или ином виде взяты из описаний архитектур IBM Blue Gene и Cray SeaStar.

В 2007 году начались работы по макетированию сети с помощью маршрутизаторов на базе ПЛИС (FPGA) [2]. В 2008 году появились первые полнофункциональные прототипы маршрутизатора (М2) на базе ПЛИС Xilinx Virtex4, с использованием которых был собран макет сети из шести узлов, соединенных в тор 3×2. Данный макет использовался для отладки базовой функциональности маршрутизатора, отработки отказоустойчивой передачи данных. Параллельно были написаны и отлажены начальные варианты драйвера и библиотеки нижнего уровня, портирована библиотека Cray Shmem и обеспечена поддержка MPI [2]. В сентябре 2010 года был запущен макет с прототипами маршрутизатора третьего поколения (М3), состоящий из девяти узлов, соединенных в двухмерный тор 3×3. В 2012 году был создан двухузловой макет для отладки высокоскоростных каналов передачи данных с пропускной способностью 12x 6,25 Гбит/с. В 2013 году появилось пер-

вое поколение маршрутизаторов сети «Ангара» на базе СБИС (рис. 2, рис. 3). В настоящий момент продолжается наладка и тестирование этих маршрутизаторов.

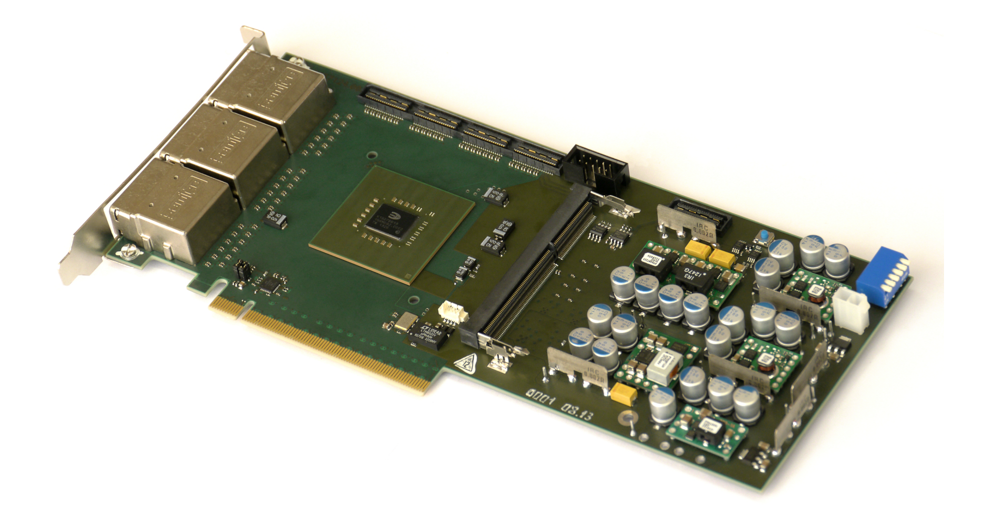


Рисунок 2. Сетевой адаптера «Ангара» на базе СБИС ЕС8430

Сравнительные характеристики сети «Ангара» с зарубежными решениями приведены в таблице 1.

Таблица 1: Сравнительные характеристики сети «Ангара» с зарубежными решениями

Характеристика	М3 (ПЛИС)	Ангара (СБИС)	InfiniBand FDR 4x	IBM Blue Gene/Q	Cray XK7
Топология сети	2D-тор	4D-тор	fat tree	5D-тор	3D-тор
ПС с процессором, ГБ/с	2	8	8	~20	9,6
ПС линка, ГБ/с	0,625	7,5	6,8	2	9,375
Агрегатная ПС линков, ГБ/с	5	120	—	40	186
Задержка между соседними узлами, мкс	2,5	1,0	1,0	$<1{,}0$	1,4

В ходе работы решались вопросы разработки методов надёжной передачи пакетов, маршрутизации и арбитража. На сетевом уровне необходимо было гарантировать, что все пакеты будут доставлены верным адресатам, никакой пакет не потеряется и не размножится, а в сети в результате работы многих узлов не возникнет тупиковых ситуаций (например, взаимных блокировок — deadlock). Для гарантии надёжной передачи по каналу связи был разработан протокол канального уровня, в рамках которого происходила нумерация пакетов и подсчёт для каждого контрольных сумм. Для предотвращения взаимных блокировок была выбрана комбинация двух методов: правило порядка направления и «правило пузырька» (bubble-rule) [5]. Отдельно потребовалось ввести дополнительный виртуальный канал для ответов на чтения, чтобы предотвратить возникновение логических взаимных блокировок, возникающих из-за взаимозависимости запросов и ответов на чтения.

Для балансировки нагрузки была добавлена адаптивная маршрутизация, для поддержки эффективной работы с сетевым адаптером многоядерных процессоров были введены несколько инжекционных конвейеров.

Взаимодействие вычислительного узла, т. е. кода, исполняемого на центральном процессоре, с маршрутизатором осуществляется путем записи данных по адресам памяти, которые отображены на адреса ресурсных регионов маршрутизатора (memory-mapped input/output). Это позволяет приложению взаимодействовать с маршрутизатором без участия ядра ОС, что снижает накладные расходы при отправке пакетов, поскольку переключение в контекст ядра и обратно занимает существенное время, в сравнении с временем отправки пакета. Для увеличения производительности было принято решение использовать аппаратно поддерживаемый механизм write-combining. Основная идея его заключается в том, чтобы временно кэшировать выданные процессором записи

во временном буфере (write combining buffer), чтобы выдать их несколько позже, вместе с последующими записями, снижая тем самым накладные расходы на выполнение отдельных операций. Поскольку запись в инжекционные буферы выполняется последовательно, write combining работает наиболее эффективным образом, что позволяет значительно повысить (до 5—20 раз) темп работы с устройством.

Так как помимо передач точка-точка значительную долю коммуникаций занимают коллективные операции (например, один узел рассылает данные группе узлов), потребовалась оптимизация выполнения таких коллективных операций, что позволило существенно повысить производительность сети на многих задачах. Коллективные операции реализуются на базе основной сети с топологией многомерный тор, при этом используются отдельные виртуальные каналы, образующие виртуальную подсеть с топологией дерево [6]. В дереве задаётся корень, относительно которого вводятся два возможных направления движения по дереву: от корня и к корню. Каждому из направлений соответствует свой виртуальный канал. Узлы, из которых движение от корня больше невозможно, называются листьями. Дерево строится с учётом порядка измерений (для предотвращения дедлоков). Для построения дерева могут использоваться вспомогательные транзитные узлы — они логически не принадлежат дереву, но нужны для его связности (в данных узлах процессоры не посылают и не получают данных).

Для достижения большей эффективности было принято решение исключить из рассмотрения случай, когда две разные задачи используют пересекающиеся группы узлов, таким образом каждый узел может относиться только к одной вычислительной задаче. Это позволило исключить накладные расходы, связанные с использованием виртуальной памяти, избежать интерференции задач, упростить архитектуру маршрутизатора за счет отсутствия необходимости в полноценном MMU и избежать всех связанных с его работой коммуникационных задержек, упростить модель безопасности сети, исключив из нее обеспечение безопасности процессов различных задач на одном узле. Принятое решение не повлияло на функциональность сети, поскольку она предназначена в первую очередь для задач большого размера. Аналогичное решение было принято в IBM Blue Gene, с той разницей, что там ограничение на единственность задачи вводится для раздела.

Основным режимом программирования для сети «Ангара» является совместное использование MPI, OpenMP и Shmem. Также поддерживаются GASNet, UPC, ARMCI, Charm++.

Выпуск СБИС потребовал решения большого числа новых задач. Хотя логика маршрутизатора к тому моменту была вполне отлажена и все базовые операции корректно работали, переход от ПЛИС к СБИС потребовал значительных усилий всего коллектива.

На начальном этапе подготовки к выпуску СБИС был проведён анализ существующих технологических возможностей. В первую очередь необходимо было выбрать технологическую норму. Выбор производился с точки зрения возможностей доступных на той или иной технологической норме различных IP-блоков (Intellectual Property Blocks) — готовых блоков, реализующих определённую функциональность. Применительно к сети «Ангара», рассматривались IP-блоки для линков, интерфейсов PCI Express, DDR 2/3 SDRAM. Также учитывался требуемый объём затрат на реализацию (дизайн топологии и изготовление) и перспективы изготовления СБИС на отечественных фабриках. Как результат, была выбрана технологическая норма 65 нм, которая, с одной стороны, была достаточно проверенной и распространённой, с другой — IP-блоки, доступные для данной технологической нормы, позволяли достигнуть исходной цели создания конкурентоспособной по производительности сети.

Немаловажным являлся вопрос проработки приоритетного варианта исполнения. Изначально предполагалось в качестве хост-интерфейса сети использовать HyperTransport, как имеющий более низкую коммуникационную задержку в сравнении с PCI Express и поддерживаемый процессорами производства компании AMD. Однако отказ AMD от поддержки не кэш-когерентного интерфейса HyperTransport и неясность дальнейших перспектив данного интерфейса, а также внедрение PCI Express root complex как составной части кристалла в процессорах компании Intel начиная с архитектуры Sandy Bridge, повлияло на данный выбор, склонив его в сторону адаптера в форм-факторе платы расширения PCI Express, как более универсального и сравнимого по эффективности решения. Вторым приоритетный форм-фактором являлась мезонинная плата для вычислительной платформы «Ангара».

Проработка форм-фактора карты расширения PCI Express привела к вопросу выбора разъёмов и кабелей. В результате рыночных исследований были выбраны разъёмы и кабели, разработанные Гейдельбергским университетом и производимые компанией Samtec [7], как обладающие уникальной плотностью упаковки (по шесть 12-лейновых разъёмов на одной планке расширения),

недоступной в иных решениях (рис. 3).



Рисунок 3. Сетевой адаптер «Ангара» на базе СБИС ЕС8430, разъёмы SAMTEC

В целях достижения целевого уровня производительности в виду предполагавшихся ограничений по частоте работы СБИС при использовании выбранного технологического процесса была удвоена ширина всех внутренних шин, что позволило достичь требуемой производительности при вдвое меньшей частоте (500 МГц при ширине 128 бит вместо 1 ГГц при ширине 64 бита). Данное фундаментальное изменение повлекло, в свою очередь, ряд изменений в архитектуре маршрутизатора: были переработаны форматы пакетов с учётом увеличившейся гранулярности флитов (16 байт вместо 8); был произведён перерасчёт необходимых размеров буферов. Одновременно с этим была добавлена поддержка адресации больших объёмов памяти: до 1 ТБ вместо 4 ГБ.

Однако основной объём работ был выполнен в рамках подготовки RTL-дизайна маршрутизатора к исполнению в СБИС. Была добавлена полноценная поддержка исключительных ситуаций, счётчиков производительности, взаимодействия с сервисным процессором, поддержка функций отладки и конфигурирования.

Механизм исключительных ситуаций позволяет путём посылки MSI[1] оповещать хост о возникновении той или иной внештатной ситуации. Необходимость генерации MSI при возникновении исключительной ситуации контролируется маской, задаваемой в адаптере сети «Ангара» посредством модификации его регистров. Также доступен механизм считывания дополнительной информации об исключительных ситуациях через интерфейс регистров адаптера, при этом для ряда ситуаций, считающихся редкими, доступна информация только о последней исключительной ситуации, для остальных же по возможности используются аккумулирующие счётчики.

Добавление счётчиков производительности было сделано в целях предоставления возможностей выявления узких мест, профилирования и оптимизации работы маршрутизатора программным обеспечением, так как в СБИС, в отличие от ПЛИС, установка подобных счётчиков путём смены прошивки по очевидным причинам невозможна. Всего в различные блоки маршрутизатора добавлено несколько сотен счётчиков, некоторые из которых предназначены для учёта только программно маркированного трафика (в дополнение к счётчикам, учитывающим весь проходящий трафик).

Была также добавлена возможность считывания конфигурации отдельных внутренних блоков и IP-блоков из Flash-памяти. Такая возможность не требовалась в ПЛИС, так как подобные вопросы решались перегенерацией прошивки ПЛИС, для СБИС же был проведён анализ по выявлению параметров конфигурации, которые необходимо настраивать на начальном этапе, и был добавлен интерфейс для их конфигурации. При невозможности настройки посредством Flash для конфигурационных параметров выбраны значений по умолчанию, некоторые из которых управляются через внешние выводы кристалла.

На этапе согласования контракта с подрядчиком производился подбор конкретных IP-блоков. При этом, помимо доступности того или иного IP-блока, необходимо было учесть доступность различных IP-блоков для выбранной толщины диэлектрика, измеряемой в напряжении пробоя, что не учитывалось при анализе на начальном этапе. При выборе между поставщиками некоторых IP-блоков принималось в расчёт наличие опыта у подрядчика по интеграции данного IP-блока.

[1] Message Signalling Interrupt, вид транзакции PCI Express, возбуждающий прерывание на хост-системе.

Из-за увеличения объёма и количества элементов SRAM в СБИС по сравнению с ПЛИС примерно в 10 раз, возникла острая необходимость добавления защиты памяти — ECC для больших элементов и битов чётности для маленьких. В процессе подбора возможного поставщика SRAM было обнаружено, что в случае использования ECC для защиты памятей существенная часть экземпляров не укладывалась во временные ограничения в связи с тем, что на вычисление ECC требовалось существенное время. Как следствие, это привело к необходимости внесения ряда изменений в дизайн — расслоения (banking) экземпляров SRAM и ослабления временных ограничений (добавления многотактовых врменных ограничений — multicycle) в тех местах, где этого было критично, например, в FIFO-буферах.

Одной из важных составляющих дизайна СБИС является DFT — Design For Testability — набор техник, позволяющих провести тестирование работоспособности чипа с точки зрения отсутствия проблем, возникающих на этапе его изготовления: замкнувших или, наоборот, разомкнутых сигналов, проблем с чтением/записью регистров, экземпляров SRAM, и так далее. Работы по внедрению структур DFT выполнялись на стороне контрагента по подготовке дизайна СБИС к выпуску. В рамках данного дизайна это привело к двум дополнительным работам. Первая задача являлась достаточно рядовой — необходимо было выполнить подготовку дизайна к внедрению структур DFT. Вторая же заключалась в необходимости интеграции разработанного механизма отладки с инфраструктурой DFT, что потребовало взаимодействие команды разработчиков ОАО «НИЦЭВТ» и группы DFT контрагента — написание варианта спецификации данного механизма для предоставления его группе DFT, проработка изменений в процессе внедрения структур DFT и самих структурах DFT в целях возможности сосуществования этих двух механизмов (механизм отладки во многом использовал те же структуры, что внедрял механизм DFT, но использование этих структур в рамках DFT в так называемом functional mode, штатном режиме работы СБИС после прохождения начального тестирования, не предполагалось, в связи с чем в схематике DFT, в основном в управляющей части, изначально имелся ряд решений, делавших подобное использование невозможным).

Одним из существенных отличий дизайна для СБИС является необходимость проработки дерева тактирования (clock tree), в то время как в случае ПЛИС оно уже имеется как часть кристалла ПЛИС. Помимо этого, необходимо было разработать схему тактирования, которая, с одной стороны, учитовала специфику тактирования отдельных IP-блоков, с другой — была достаточно гибкой и управляемой (разные варианты получения базового тактового сигнала, используемого логикой маршрутизатора и контроллером PCI Express), с третьей — учитывала требования, предъявляемые DFT и позволяла реализовать возможности механизмов отладки. Не всё из перечисленного удалось успешно реализовать; в частности, пришлось пожертвовать одной из планировавшихся техник отладки — мгновенной (в течение одного такта после срабатывания триггера) приостановкой тактирования логики маршрутизатора: дерево тактирования оказалось настолько большим, что между моментом генерации тактового импульса и его дохождением до тактируемой им логикой проходило несколько тактов; большим, в свою очередь, оно оказалось из-за углового расположения тактового генератора, которое, в свою очередь, было обусловлено организацией обеспечения питания.

Отдельно стоит упомянуть про процесс разводки подложки (substrate) СБИС. Выбор платы расширения PCI Express в качестве одного из основных вариантов исполнения накладывает ограничения на количество слоёв печатной платы вследствие технологических возможностей производства ОАО «НИЦЭВТ». В целях минимизации числа требуемых слоёв для вывода контактов со СБИС к расположению выходных контактов на подложке предъявлялся ряд определённых требований, в первую очередь, касающихся расположения выводов дифференциальных пар линков и PCI Express (их порядка и взаимного расположения), что, в свою очередь, породило сложности в процессе разводки подложки, так как максимально допустимое количество слоёв подложки было, в свою очередь, обусловлено подписанным контрактом; помимо этого были чисто технические трудности с нахождением пространства для размещения переходных отверстий.

Много нетривиальных задач, связанных с подготовкой СБИС, встало перед группой моделирования и верификации ОАО «НИЦЭВТ». Появление дерева тактирования повлекло за собой необходимость моделирования нетлиста (вместо RTL), причём с учётом задержек (netlist timing simulation). Самое дерево тактирования, привнося с собой большое количество буферов задержки, существенно увеличивало объём модели. Помимо этого, необходимость включать большое количество аналоговых моделей (только SerDes'ов линков и PCI Express суммарно насчитывается 112 экземпляров; интерфейс DDR SDRAM имеет сравнимый объём) также весьма существенно сказалось на скорости моделирования и объёмах требуемой для данного процесса памяти. Был ряд проблем

и с самими моделями IP-блоков — в процессе их использования иногда приходилось сталкиваться с ошибками в моделях (особенно тяжело диагностируемыми были проблемы, приводившие к зацикливанию процесса моделирования); некоторые модели не могли быть проаннотированы (что также вызвало дополнительные сложности при использовании их в моделировании с учётом временных задержек). Одной из важных задач, стоявших перед группой моделирования и верификации, являлась необходимость моделирования взаимодействия нескольких маршрутизаторов, объединённых в сеть. В связи с существенным увеличением объёма моделей, моделирование не могло, как раньше, выполняться в рамках одного узла (процесс моделирования занимал слишком много времени и требовал большое количество оперативной памяти, существенно сужая круг доступных вычислительных ресурсов для запуска моделирования на них), в связи с этим тестовое окружение было доработано: была добавлена поддержка взаимодействия процессов моделей по MPI, что позволило осуществлять запуски на многоузловых вычислительных системах и существенно сократить время, требуемое на получения результатов данного вида моделирования. Отдельно можно отметить, что интенсивное использование моделирования в процессе подготовки СБИС позволило выявить некоторые потенциальные проблемы, которые не были выявлены статическим временным анализом (static timing analysis, STA).

Помимо подготовки дизайна на уровне RTL для её передачи контрагенту, был проведён существенный объём работ, касающийся процесса синтеза нетлиста пригодного для использования контрагентом в процессе разводки топологии и внедрения DFT. В частности, были согласованы особенности именования отдельных элементов и сигналов, требуемая иерархия блоков для удобства работы контрагента с ними (RTL-вариант дизайна имеет довольно глубокую иерархию, которая излишне ограничивает и затрудняет процесс подготовки топологии, в связи с чем потребовалось выполнять упрощение иерархии в процессе синтеза; свои требования были к предоставляемому нетлисту у группы DFT).

Кроме технических сложностей, в проекте такого масштаба были и чисто организационные, связанные с естественной необходимостью тесного взаимодействия разработчиков из нескольких крупных компаний из разных стран мира, между которыми имелись языковые, понятийные, юридические, субординационные и другие подобного рода барьеры. При взаимодействии с крупными международными контрагентами оказалось крайне трудно бороться с тенденцией к сведению всех вопросов лишь к стандартным и наиболее легко реализуемым решениям, слабо учитывающим специфику конкретного проекта.

Заключение

Кристалл СБИС ЕС8430 изготовлен на фабрике TSMC с использованием технологических норм 65 нм, имеет размеры 13,0×10,5 мм, содержит 180 миллионов транзисторов; корпусировка FCBGA (flip-chip ball grid array), 1521 вывод в виде массива 39×39 контактов с шагом 1 мм, подложка имеет размеры 40×40 мм. Плата сетевого адаптера изготавливается на собственном производстве в ОАО «НИЦЭВТ». СБИС работает на частоте 250/500 МГц (в зависимости от используемой скорости PCI Express) и потребляет 36 Вт энергии. Плата маршрутизатора позволяет подключить до 6 линков (до 8 с платой расширения) пропускной способностью 75 Гбит/с каждый (кодирование 8b10b). Взаимодействие адаптера с вычислительным узлом осуществляется через PCI Express 2.0 x16 (80 Гбит/с, кодирование 8b10b).

Продвижение сети «Ангара» на рынок планируется осуществлять в двух вариантах: как отдельную коммерческую сеть в виде плат PCI Express для кластерных систем с коммерчески доступными серверными узлами, и как интегрированный компонент в составе разрабатываемой в ОАО «НИЦЭВТ» в рамках проекта «Ангара» вычислительной платформы, что позволит объединить до 32 тысяч узлов в составе суперкомпьютера транспетафлопсного уровня производительности.

Параллельно с выпуском СБИС первого поколения продолжается дальнейшая разработка и оптимизация архитектуры сети «Ангара», готовится макет М4 (на базе ПЛИС Virtex 7). Опыт эксплуатации предыдущих макетов и кластера с маршрутизаторами на базе СБИС является основой для разработки принципов работы сети «Ангара» второго поколения. Основные доработки будут направлены на поддержку большего числа топологий, повышение безопасности выполнения прикладных задач на узлах, добавление аппаратной поддержки атомарных операций с возвратом значений, поддержки технологии GPU Direct, оптимизацию RDMA-операций и поддержки большого числа тредов/процессов на узле.

Список литературы

[1] Д. В. Макагон, Е. Л. Сыромятников *«Сети для суперкомпьютеров»*, http://www.osp.ru/os/2011/07/13010500/

[2] А. А. Корж, Д. В. Макагон, И. А. Жабин, Е. Л. Сыромятников и др. *«Отечественная коммуникационная сеть 3D-тор с поддержкой глобально адресуемой памяти для суперкомпьютеров транспетафлопсного уровня производительности»*, Параллельные вычислительные технологии (ПаВТ'2010): Труды международной научной конференции (Уфа, 29 марта — 2 апреля 2010 г.), http://omega.sp.susu.ac.ru/books/conference/PaVT2010/full/134.pdf — Челябинск: Издательский центр ЮУрГУ, 2010. С. 227—237.

[3] J. Dongarra *«Visit to National University for Defense Technology Changsha»*, China. June 3, 2013.

[4] William Dally and Brian Towles, *Principles and Practices of Interconnection Networks*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[5] Jose Duato, Sudhakar Yalamanchili, and Ni Lionel. *Interconnection Networks: An Engineering Approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.

[6] Е. Л. Сыромятников, Д. В. Макагон, С. И. Парута, А. А. Румянцев *«Реализация аппаратной поддержки коллективных операций в маршрутизаторе высокоскоростной коммуникационной сети с топологией "многомерный тор"»*, Перспективные направления развития средств вычислительной техники: Труды ОАО «Научно-исследовательский центр электронной вычислительной техники», — Москва: Издательство «Радиотехника», 2012. С. 11—15.

[7] *HDI6 - 0,635 mm Eye Speed® HD High Speed High Density Receptacle*, http://www.samtec.com/technical-specifications/default.aspx?seriesMaster=HDI6